Eur J Epidemiol (2010) 25:875–883
DOI 10.1007/s10654-010-9518-5

DEVELOPMENTAL EPIDEMIOLOGY

Early life patterns of common infection: a latent class analysis

Sarah J. Hepworth · Graham R. Law · Debbie A. Lawlor · Patricia A. McKinney

Received: 22 March 2010 / Accepted: 12 October 2010 / Published online: 26 October 2010

Abstract Early life infection has been implicated in the aetiology of many chronic diseases, most often through proxy measures. Data on ten infectious symptoms were collected by parental questionnaire when children were 6 months old as part of the Avon Longitudinal Study of Parents and Children, United Kingdom. A latent class analysis was used to identify patterns of infection and their relationship to five factors commonly used as proxies: sex, other children in the home, maternal smoking, breastfeeding and maternal education. A total of 10,032 singleton children were included in the analysis. Five classes were identified with differing infectious disease patterns and children were assigned to the class for which they had a highest probability of membership based on their infectious symptom profile: 'general infection' (n = 1,252, 12.5%), 'gastrointestinal' (n = 1,902, 19.0%), 'mild respiratory' (n = 3,560, 35.5%), 'colds/ear ache' (n = 462, 4.6%) and 'healthy' (n = 2,856, 28.5%). Females had a reduced risk of being in all infectious classes, other children in the home were associated with an increased risk of being in the 'general infection', 'mild respiratory' or 'colds/ear ache' class. Breastfeeding reduced the risk of being in the 'general infection' and 'gastrointestinal' classes whereas maternal smoking increased the risk of membership. Higher maternal education was associated with an increased risk of being in the 'mild respiratory' group. Other children in the home had the greatest association with infectious class membership. Latent class analysis provided a flexible method of investigating the relationship between multiple symptoms and demographic and lifestyle factors.

Keywords ALSPAC · Latent class analysis · Infection · Proxies · Hygiene hypothesis

S. J. Hepworth (✉) · P. A. McKinney
Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Room 8.49, Paediatric Epidemiology Group, Worsley Building, Clarendon Way, Leeds LS2 9JT, UK
e-mail: s.j.hepworth@leeds.ac.uk

G. R. Law
Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Room 8.001, Biostatistics Unit, Worsley Building, Clarendon Way, Leeds LS2 9JT, UK

D. A. Lawlor
Division of Epidemiology, Department of Social Medicine, University of Bristol, Bristol, UK

D. A. Lawlor
Medical Research Council Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, UK

Introduction

Early life common infection has been of heightened interest to researchers due to the proposed 'hygiene hypothesis' [1–3]. This suggests that lack of exposure to common pathogens during infancy in modern times in developed countries may have a long lasting effect on the immune system [4]. Lack of exposure is suggested to increase the risk of future development of chronic disease with underlying infectious aetiology including allergies, asthma [5], Type 1 diabetes [6] and childhood acute lymphoblastic leukaemia [7].

Studies have found that at birth many facets of immune response are immature and have a different profile when compared to adults [8, 9]. During the neonatal period and beyond a child's immune system continues to mature towards levels of response seen in adulthood and the first

2 years are seen as a particularly important period in this development [10–13]. Environmental exposures including to infectious pathogens during infancy are likely to affect the immune system though the mechanisms involved are still to be fully understood [4].

In studying these hypotheses, data on infections in infancy can be difficult to collect as chronic diseases are usually diagnosed later in life. Proxies for levels of common infection such as number of siblings or breastfeeding have been used instead [14, 15]. An alternative to assessing these hypotheses would be to use medical records in regions where record linkage is possible. However, such records include only episodes of infection requiring a visit to a doctor and this will only represent a small percentage of overall common infancy infectious episodes. In a recent Swedish study it was shown for 18 month old children that only 13% of infectious episodes led to contact with a healthcare provider [16]. The decision by parents to take their child to the doctor can be affected by individual factors unrelated to illness including whether the child is first born and parental education and concern levels about infectious illness, which may also be related to later outcomes and hence confound any apparent associations [17, 18].

More information is required on patterns of common infection amongst young children and how they relate to proxies used to represent common infection levels in studies of later chronic disease. This may allow us to gain more insight into how different proxies may represent different patterns of exposure to common infectious diseases of different types.

Latent class analysis is a statistical method that can be used to identify groups of individuals with similar patterns of responses [19]. The aim of this analysis was to identify different classes of reported infectious symptom profiles during the first 6 months of life using data from the Avon Longitudinal Study of Parents and Children (ALSPAC) [20]. The associations between demographic and lifestyle factors and the class membership were also examined to gain insight into what aspects of infectious symptoms are reflected by proxies for common infection used in studies of chronic disease.

Methods

Participants

The ALSPAC study recruited women resident in the old county of Avon in South West England, with an expected date of delivery 1st April 1991 to 31st December 1992. A total of 14,541 pregnant women were recruited with a resulting 14,062 live births. In this analysis only singleton births were considered for inclusion in the analysis ($n = 13,678$) where the 6 month questionnaire had been completed including the questions on infectious symptoms ($n = 11,198$). Ethical approval for the study was obtained from the ALSPAC Law and Ethics Committee and local NHS research council.

Data used in the analysis

Questions were completed by the mother or main caregiver for the child when they were 6 months old and included the question 'has your child had any of the following?'. For which 13 symptoms and signs which may have an infectious cause were listed: diarrhoea, blood in the stools, vomiting, cough, high temperature, snuffles/cold, earache, ear discharge (pus not wax), convulsions/fits, colic, rash, wheezing and breathlessness. Three responses were available: 'Yes, and did not see a doctor', 'Yes, and did see a doctor' and 'No', these have been converted in this analysis to a dichotomous yes/no variable of symptom presence such that the first two responses were combined to create the 'yes' response. To allow investigation of symptom clusters three symptoms: convulsions/fits, colic and rash were a priori excluded from the analysis due to their lack of specificity for an infectious cause.

Alongside this information several variables related to infection were extracted from the ALSPAC data-set. These were chosen based on their previous use as proxies for common infection in previous studies of the aetiology of chronic disease or due to findings of association with infection levels. Sex of the child was included due to the previously found excess of infectious disease hospitalization and mortality in males especially in relation to respiratory illnesses and this may also influence manifestation of less severe infectious disease symptoms [21]. Older children or siblings in the home is commonly used as a proxy for infection in studies of cancer risk [7, 22, 23] and allergic disease [14, 24]. In this case the questionnaire was related to older children within the home and was collected at the same time as the infectious symptom data. Smoking in the home has also been previously associated with higher levels of respiratory infection in children [25]. Maternal smoking collected in a questionnaire completed by the mother when the child was 8 months old was chosen as the representative variable for environmental tobacco smoke exposure as smoking data for the mother or main caregiver's partner was missing for a significant minority of children. This was the earliest time point after the birth information on smoking was collected. Breastfeeding was considered as a proxy using the same definition as an earlier ALSPAC analysis where information on whether the child had ever been breastfed based on questionnaire information provided by the mother or main caregiver [26]. Breastfeeding has previously been used as a proxy for infection levels in studies of leukaemia and lymphoma [27, 28] as well as allergic disease

[29]. Maternal education was used as an indicator of socio-demographic status and was collected by questionnaire during pregnancy.

To confirm the results found for maternal smoking a further model was fitted including partner's smoking instead of maternal smoking to investigate if the direction of results was the same.

Statistical methods

Latent class analysis was used to explore the number of distinct classes of children displaying different patterns of common infection in the first 6 months of life. Latent class analysis is a statistical method that allows the classification of individuals into groups based on conditional probabilities [30]. It is part of the larger family of latent variable modelling which is based on the assumption of underlying variables that cannot be measured directly only indirectly through observed data [31]. Within each class, individuals will have a similar pattern of responses to categorical or continuous variables [19].

To decide on the correct number of classes several indices of fit can be used. These include Akaike's information criterion, the Bayesian information criterion and also a likelihood ratio test using bootstrapping [32]. Alongside the use of statistical tests the class structure of the different models were compared, as an important aspect of this type of analysis is the ability to meaningfully interpret the classes [33]. An appendix to this paper provides further information on the methods and the results for the models with 1–8 latent classes when fitted to the data. For this method local independence was assumed, this means that within a class the presence of a symptom is independent of another symptoms presence as all correlation between the variables is explained through the class structure.

The interpretation of classes is carried out by comparing the probability of each symptom in each class to try to find which features distinguish each group from the others. The relative probability of the symptom in each class needs to be considered as well as the overall probability [33]. This allows labels and characteristics to be assigned to the different classes.

In this analysis there were two stages to fitting the final chosen model. In the first stage a model just containing infectious symptoms was used to find the optimal number of classes, in the second stage the model with the optimal number of classes was used in a single step regression including all covariates to predict class membership and the influence of covariates on this.

The estimated probability of membership of each class is calculated for each child based on the combination of recorded symptom responses and covariates which will sum to one for each child. The relative size of each class was investigated using modal assignment where children are assigned to the class for which they have the highest probability of membership [34]. Analysis was carried out using the M-Plus statistical package [35]. Participants with missing values for one or more covariates were excluded from the second stage of the analysis.

Conditional probabilities are presented from this final model, which is the probability of the symptom being present conditional on membership of that class, with the optimum number of classes and including the covariates (sex, older children, maternal smoking, breastfeeding, and maternal education). Results for the effect of the covariates are presented as odds ratios comparing the odds of being in the class under investigation compared to the comparison class. This is equivalent to carrying out a multinomial logistic regression across the separate classes using one as a reference group [36].

Results

Of the ten infectious symptoms included in the latent class analysis the most common three symptoms were cold (87%), cough (64%) and high temperature (39%) with ear discharge (the most uncommon) only reported in 2.8% of children (Table 1). The most common combination of symptoms reported was cough and cold present in 12.6% of children. Descriptive statistics of the demographic factors used as covariates in the analysis are given in Table 2.

Models with between one and eight classes were considered including the ten symptoms postulated to be most strongly related to infection. The final chosen model contained five classes after consideration of the statistical measures of fit and the interpretability of the results (see appendix). The predicted probability for each symptom within each latent class is given in Table 3. Class assignment and structure were similar when all participants were included compared to when those only with all covariate data were included in the five class model (data not shown).

Table 1 Number of children in the Avon longitudinal study of parents and children (n=11,198) for which each of the ten infection symptoms were reported when the child was 6 months old

Symptom	Number	Percentage
Diarrhoea	3,730	33.3
Blood in the stools	442	3.9
Vomiting	3,499	31.2
Cough	7,262	64.9
High temperature	4,392	39.2
Snuffles/Cold	9,810	87.6
Ear ache	1,135	10.1
Ear discharge	316	2.8
Wheezing	2,408	21.5
Breathlessness	688	6.1

Table 2 Demographic factors amongst the 6 months old children in the Avon longitudinal study of parents and children (n=11,198)

Variable		n	%
Sex	Female	5,769	51.5
	Male	5,429	48.5
	Missing	0	0.0
Older children	Yes	6,218	55.5
	No	4,908	43.8
	Missing	72	0.6
Maternal smoker	Yes	2,406	21.5
	No	8,013	71.6
	Missing	779	7.0
Breastfeeding	Yes	8,382	74.9
	No	2,812	25.1
	Missing	4	<0.1
Maternal education	High school or less	6,729	60.1
	College qualification	4,008	35.8
	Missing	461	4.1
Partner smoker	Yes	6,681	59.7
	No	2,097	18.7
	Missing	2,420	21.6

A total of 10,032 (89.6% of those eligible) participants were included in the final analysis examining the association of covariates with probability of class membership (Table 3), after exclusion of participants with missing covariate data (n = 1,166, 10.4%). The five classes had varying numbers of members; the first class 'general infection' (n = 1,252, 12.5%) had a high probability of all infectious symptoms and specifically a high probability of breathlessness and wheeze symptoms compared to the other classes. Children within the second 'gastrointestinal' class (n = 1,902, 19.0%) have the highest probability of sickness and diarrhoea. The third group have 'mild respiratory' symptoms with lower probabilities of wheeze or ear related symptoms than the other two groups where children had higher probability of respiratory symptoms (class 1, 4) (n = 3,560, 35.5%). The fourth 'colds/ear ache' (n = 462, 4.6%) class contained children with a similar probability of respiratory symptoms but a higher probability of ear ache and ear discharge. The final 'healthy' (n = 2,856, 28.5%) group have the lowest probability of any of the infectious symptoms.

Interpreting the effect of covariates on the model the 'healthy' class was chosen as the comparison group as they had the lowest probability for all symptoms. Table 4 gives the multivariable odds ratios for each covariate for membership of each class compared to this comparison class.

Risk of being in the 'general infection' class was associated with all factors except maternal education. The factor associated with the greatest increased odds ratio of being within this class was having older children in the household. Maternal smoking was also an indicator of increased risk. Both breastfeeding and being female was associated with decreased odds.

Being female and being breastfed were both associated with lower odds of being in the 'gastrointestinal' class. Maternal smoking was associated with increased odds for being in the 'gastrointestinal' class, whereas the presence of older children in the home and maternal education was not associated with odds of being in this class.

Factors associated with being in the 'mild respiratory' class were sex, older children and maternal education. As for previous classes being female decreased the odds of class membership whereas older children and a higher level of maternal education increased the risk, this pattern was also seen for the 'colds/ear ache' class which contained those with a similar symptom profile to this class but with a higher probability of ear ache and ear discharge.

Table 3 Conditional probability for each symptom within each of the 5 classes using the Avon longitudinal study of parents and children infectious symptom data from 6 months and including covariates in the model

Class Assigned label	1 General infection	2 Gastrointestinal	3 Mild respiratory	4 Colds/ear ache	5 Healthy
% in class*	12	20	37	5	26
Number in class	1,252	1,902	3,560	462	2,856
1.Diarrhoea	0.545	0.844	0.089	0.368	0.160
2.Blood in stools	0.074	0.065	0.022	0.050	0.029
3.Vomiting	0.565	0.624	0.186	0.381	0.120
4.Cough	0.972	0.726	0.869	0.775	0.090
5.High temp	0.707	0.482	0.341	0.773	0.170
6.Snuffles/cold	0.966	0.933	0.966	0.969	0.648
7.Ear ache	0.177	0.085	0.031	0.857	0.031
8.Ear discharge	0.055	0.008	0.006	0.306	0.007
9.Wheezing	0.916	0.102	0.187	0.106	0.019
10.Breathlessness	0.412	0.016	0.009	0.018	0.003

* Percentage according to the posterior probability totals

Table 4 Odds ratio results for membership of the latent classes compared to the 'healthy' class of children from the Avon longitudinal study of parents and children

Variable	Reference	Comparison	Class 1			Class 2			Class 3			Class 4			Class 5
			General infection			Gastrointestinal			Mild respiratory			Colds/ear ache			Healthy
			OR	95% CI		OR	95% CI		OR	95% CI		OR	95% CI		OR
Sex	Male	Female	0.48	0.40	0.57	0.83	0.71	0.98	0.70	0.61	0.80	0.71	0.54	0.92	1.00
Older children	No	Yes	3.46	2.75	4.36	1.03	0.86	1.24	1.81	1.52	2.16	2.72	1.94	3.80	1.00
Maternal smoker	No	Yes	1.59	1.31	1.93	1.23	1.01	1.49	0.84	0.70	1.01	0.99	0.71	1.37	1.00
Breast-feeding	No	Yes	0.63	0.51	0.77	0.52	0.43	0.64	1.05	0.85	1.29	1.08	0.74	1.60	1.00
Maternal education	High school or less	College qualification	0.96	0.79	1.16	0.85	0.70	1.02	1.28	1.09	1.50	1.13	0.82	1.56	1.00
Partner smoker[a]	No	Yes	1.51	1.22	1.87	1.15	0.93	1.43	0.74	0.60	0.91	0.86	0.61	1.22	1.00

[a] Added to a separate model including all other covariates except maternal smoking

The inclusion of partner's smoking showed a similar pattern to maternal smoking with the association between smoking and membership of the 'general infection' class being raised but no association with membership of the other classes except for a significantly reduced association with membership of the 'mild respiratory' class. Information on partner's smoking was missing for a further 1,549 children who were included in the main analysis.

Discussion

New ways to characterize the complexity of early life infection and associations with factors often used as proxies for common infection when investigating later health is important for research into child and adult chronic diseases. The different classes of children in our analysis define presentations of common infectious symptoms and are likely to represent those with different infectious exposure. Their relationship with factors which are commonly used as proxies for infection and other related factors gives new insight into what proxies may actually represent in terms of differences in common infectious symptoms.

Our analysis identified 5 classes to which children could be assigned with differently reported combinations of symptoms at age 6 months defined as 'general infection', 'gastrointestinal', 'mild respiratory', 'colds/ear ache' and 'healthy'. Investigation of the relationship between these and five potential proxy factors of infectious exposure (the child's sex, other children, maternal smoking, breastfeeding and maternal education) revealed several associations between the odds of being in each symptom class compared to the 'healthy' group. Female infants were associated with lower odds of being in any of the infection groups compared to the 'healthy' class. Older children in the home were associated with an increased risk of the three respiratory symptom classes: 'general infection', 'colds/ear ache' and 'mild respiratory' and this covariate had the strongest magnitudes of association of any covariates with class membership. Maternal smoking increased the odds of membership of both the 'general infection' and 'gastrointestinal' classes. Breastfeeding reduced the odds of membership of the 'general infection' and the 'gastrointestinal' class, but was not strongly associated with other infancy infectious disease classes. Higher maternal education increased the risk of being in the 'mild respiratory' infection category.

Other studies have investigated the effects of demographic and lifestyle factors selected by our study in association with individual infectious pathogens including bacterial and viral disease. A recent study of children infected with and/or hospitalized for human metaneumovirus (hMPV) and respiratory syncytial virus (RSV) in Copenhagen [37] reported having older siblings was positively associated with hMPV but not RSV infection. Smoking within the household was positively associated and breastfeeding negatively associated with RSV but not hMPV infection. RSV and hMPV are common causes of upper and lower respiratory tract infection in children under 5 years, with symptoms from the two being indistinguishable but with some differences in their epidemiology [38]. Thus the findings from this Danish study have some similarity with ours in that having older children in the home was strongly associated with all three classes that had features of upper and lower respiratory tract infection in our study. In our study maternal smoking was positively

associated and breastfeeding protective of a high probability of being in the 'general infection' class, which had wheezing and lower respiratory tract infection symptoms as important characteristics, but there was no strong evidence of these two characteristics being associated with the classes representative of upper respiratory tract or mild respiratory infection ('mild respiratory' and 'colds/ear ache' classes). These differences might be explained by the fact that our study included all symptoms as reported by the main carer including those where no medical intervention was required, whereas in the Danish study participants were those who were receiving care from the health services.

A US study of outpatient and inpatient visits for bronchiolitis identified being male, having older siblings and not initiating breastfeeding increased the risk of bronchiolitis [39]. The results from both these studies are consistent with our findings with respect to sex, older children in the home and breastfeeding and show that these factors may not only relate to low level infections but also relate to more serious manifestations requiring hospitalization or the use of medical services though there was no association with smoking or breastfeeding and membership of the mildest respiratory infection group (class 3).

Support for our findings on breastfeeding comes from a systematic review of the effects of breastfeeding on both gastrointestinal and respiratory infection which found the majority of studies reported a protective effect on gastrointestinal conditions and more severe respiratory infection seen in the 'general infection' class [40, 41].

In our study females had reduced odds of being in any of the infectious disease classes. These findings are consistent with long established findings of higher infant mortality in males, which at least in part is related to their greater risk of infant infections [42, 43], in particular respiratory infections [21], these current findings suggest the pattern of male excess is also present when considering milder infectious disease symptoms though through what mechanism these differences occur is unknown.

One of our study findings was that maternal smoking, already a well established risk factor for respiratory infection [44] may also increase the risk of gastrointestinal type symptoms in children. Positive relationships between maternal smoking and diarrhoea [45] and other gastrointestinal symptoms such as reflux and infantile colic [46] have been reported. Our study suggests that instead of increasing the risk of respiratory infection overall the risk was specific to the potentially more severe general infection where children had the highest probabilities for breathlessness and wheeze (class 1) as opposed to mild infection (class 3). Partner's smoking showed a similar pattern but the association was not as strong possibly because of the lower amount of time spent with the child. In America married mothers have been reported as spending twice as much time in child care activities compared to fathers though smoking was missing for a greater percentage of partners and this may also have had an effect on results [47]. Smoking status was also collected 2 months after the infection data and smoking status may have changed during this time though it was assumed average smoking levels would remain similar.

In this study we have shown how latent class modelling can be used to combine a number of carer reported infant infection symptoms into an interpretable set of classes with different infection patterns that it is then possible to explore with respect to key risk factors for each class. A latent class model was fitted including the infectious symptoms and explanatory variables within one combined latent class model. This was to ensure the model was classified using all available information and to allow the estimation of model parameters for their effect that took into account all the uncertainty associated with the allocation of a probability of membership for each of the five classes for each child [36]. There are some issues related to the latent class method. Much of the interpretation of classes is based on individual opinion and different characteristics can be used to describe the same class, an example within this study is the first class described as 'general infection' due to relatively high probabilities of all symptoms when compared to the other classes and this overall profile was seen as of most importance but due to the higher levels of breathlessness and wheeze in this group this could also be chosen as the distinguishing factor and a label of 'severe respiratory' or 'asthma-like symptoms' may also be appropriate but may depend on the context [33]. The choice of model with the optimal number of classes can also vary depending on the method used and can again be open to interpretation [32].

Symptoms are one observable manifestation of infection but many infections especially if they are viral may have no symptoms or may lead to very mild symptoms. The ALSPAC questionnaire, completed by the child's mother or main caregiver when the child was 6 months of age was a non-invasive method of measurement. On a large scale this approach has many advantages including repeated data-collection over time which would give insight into how patterns of infection change as children get older. However, the data can only give an approximation of the level of actual infection. In addition, perceptions of illness may differ between the individuals reporting information and it is impossible to make adjustments for this in the analysis [18]. One approach to validation of reports of infections is to compare parental/carer reports with medical records but this is compromised by differences in health seeking behaviours between individuals [48]. An ideal scenario would be to validate reports of infections against serological testing of blood samples but no reports in the literature could be found.

Symptoms were in this analysis categorised as a binary 'yes/no' variable; other categorisations could have been used including a binary variable for symptoms requiring a doctors visit or an ordinal three category variable possibly representing severity of symptoms as symptoms requiring a doctors visit could be assumed to be more serious. This may have given more information and it could be assumed symptoms requiring a doctors visit may be more accurately recalled due to a medical intervention taking place but this is difficult to quantify and having three categories would have made interpretation more complex. The data collected did not include information on number of episodes and this may be an important metric when considering infection in the context of the hygiene hypothesis. These data were not available within the ALSPAC study but would have been included if available.

Seasonality is strongly associated with patterns of infections especially during the winter months. Seasonality was not considered within our analysis as a primary aim was to identify different classes of children with different infectious experiences irrespective of a seasonal effect. The months of completion of the questionnaire were evenly distributed throughout the year and the distribution of the demographic factors remained unaffected by the month of the year the questionnaire was completed (results not shown). Other environmental and demographic factors are also likely to influence levels of common infection in early life such as hygiene levels in the home, social contacts outside the home, ethnicity and birth related factors such as birth weight, gestational age and method of delivery. In this analysis we focused on factors previously used as proxies for common infection in studies of chronic disease and to allow illustration of the methodology whilst trying to keep levels of multiple testing to a minimum.

When the hygiene hypothesis is being considered proxy measures are used as a measure of overall infectious challenge rather than a proxy only relating to a specific type of infection such as respiratory or gastrointestinal. Overall our analysis showed that having older children in the home had overall the strongest association with membership of all infectious symptom classes and as a proxy may be the most appropriate of those considered as an overall measure of infectious exposure. Though in other cases a different measure may be required where a specific sort of infection is expected to be a risk factor. This analysis has shown which proxies may be more appropriate depending on the type of infection of most interest or whether a combination of proxies may be required.

In conclusion, latent class analysis of a large dataset of reported infectious symptoms on children at 6 months of age identified five classes of children with differing patterns of infectious disease. These were found to be related to five demographic and lifestyle factors commonly used as proxies for infection. These findings will help to inform the choice of proxy measures in future analyses of early infections and later health.

Appendix: statistics for choosing the latent class model that best fits the data

There are several criteria for assessing the number of classes which appropriately describe the different subgroups of the population when fitting a latent class model. Because the models with more classes aren't nested within previous models differences in log likelihood using the likelihood ratio test can't be used [49]. One way is to compare the percentage change in the likelihood ratio chi-square statistic from each model for see when there is not much discernable difference by adding an extra class. A further model is entropy. When each latent class model is fitted each individual is assigned a probability of belonging to each class. They are then usually assigned to the class for which they have the highest probability of membership (model assignment). One way to measure how well the participants are assigned to classes is the overall entropy. Values of entropy fall between 0 and 1 and the closer to 1 the better the classification value of the model.

Other methods include using Information Criterion which measure a combination of model fit with penalization for additional classes in the model. The most commonly used in latent class analysis are Akaike's, Bayesian and Bayesian adjusted for sample size, the model with the lowest values of the statistic is seen as having the best fit to the data [32]. Another measure of model fit is the parametric bootstrapped likelihood ratio test [50], this can be used to check if a model with one less class in the model is a significantly worse fit than the chosen number of classes. Another important aspect of assessment of model fit is also whether the model is supported by related theory and whether the classes can be interpreted as meaningful [33].

Between 1 and 8 classes were fitted to the data and the results examined (Table 5). A scree plot was chosen to

Table 5 Model fit statistics from the ALSPAC symptom data for models with between 1 and 8 latent classes

Model	L statistic	Entropy	% reduction from H_0	AIC	BIC	SS adjusted BIC
1	7,530			96,871	96,944	96,913
2	3,093	0.560	58.9	92,220	92,374	92,307
3	2,267	0.606	69.9	91,415	91,650	91,548
4	1,461	0.601	80.6	90,631	90,946	90,810
5	985	0.655	86.9	90,177	90,573	90,401
6	767	0.643	89.8	89,991	90,467	90,260
7	632	0.627	91.6	89,868	90,425	90,183
8	558	0.653	92.6	89,826	90,463	90,186

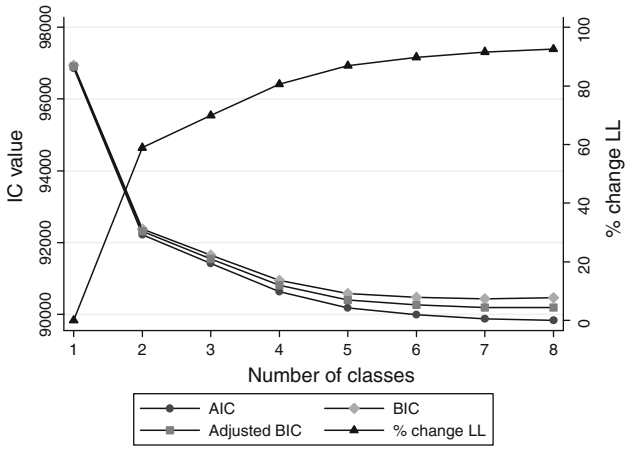

Fig. 1 Model fit statistics for the ALSPAC symptom data for between 1 and 8 classes plotted as a scree plot

allow comparison of the change in the statistics with each model (Fig. 1). As can be seen the BIC statistic suggests the 7 class model as the best fit but further examination of the 5 and 6 class models showed the increase in fit was only marginal with the addition of two extra classes. Examination of the probabilities from the models suggested that the 5 class model had the easiest interpretability when considering general classes of infectious symptoms and was chosen as the final model. The bootstrapped likelihood ratio statistic was highly significant ($P < 0.001$) suggesting that the 5 class model was a significantly better fit to the data than the 4 class model.

Acknowledgments We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. This publication is the work of the authors and SH, GL, PMcK and DL will serve as guarantors for the contents of this paper. SH is supported by the EU Integrated Project NewGeneris, 6th Framework Programme, Priority 5: Food Quality and Safety (Contract no. FOOD-CT-2005-016320). NewGeneris is the acronym of the project 'Newborns and Genotoxic exposure risks' http://www.newgeneris.org. The UK Medical Research Council, the Wellcome Trust and the University of Bristol provide core support for ALSPAC.

References

1. Garn H, Renz H. Epidemiological and immunological evidence for the hygiene hypothesis. Immunobiology. 2007;212:441–52.
2. Wills-Karp M, Santeliz J, Karp CL. The germless theory of allergic disease: revisiting the hygiene hypothesis. Nat Rev Immunol. 2001;1:69–75.
3. Yazdanbakhsh M, Kremsner PG, van Ree R. Allergy, parasites, and the hygiene hypothesis. Science. 2002;296:490–4.
4. West LJ. Defining critical windows in the development of the human immune system. Hum Exp Toxicol. 2002;21:499–505.
5. von Mutius E. Allergies, infections and the hygiene hypothesis–the epidemiological evidence. Immunobiology. 2007;212:433–9.
6. Cooke A. Review series on helminths, immune modulation and the hygiene hypothesis: how might infection modulate the onset of type 1 diabetes? Immunology. 2009;126:12–7.
7. Law GR. Host, family and community proxies for infections potentially associated with leukaemia. Radiat Prot Dosimetry. 2008;132:267–72.
8. Marodi L. Down-regulation of Th1 responses in human neonates. Clin Exp Immunol. 2002;128:1–2.
9. Holt PG, Jones CA. The development of the immune system during pregnancy and early life. Allergy. 2000;55:688–97.
10. Upham JW, Lee PT, Holt BJ, Heaton T, Prescott SL, Sharp MJ, Sly PD, Holt PG. Development of interleukin-12-producing capacity throughout childhood. Infect Immun. 2002;70:6583–8.
11. Clerici M, DePalma L, Roilides E, Baker R, Shearer GM. Analysis of T helper and antigen-presenting cell functions in cord blood and peripheral blood leukocytes from healthy children of different ages. J Clin Invest. 1993;91:2829–36.
12. Prescott SL, Macaubas C, Smallacombe T, Holt BJ, Sly PD, Holt PG. Development of allergen-specific T-cell memory in atopic and normal children. Lancet. 1999;353:196–200.
13. Savelkoul HF, Neijens HJ. Immune responses during allergic sensitization and the development of atopy. Allergy. 2000;55:989–97.
14. Bernsen RM, de Jongste JC, van der Wouden JC. Birth order and sibship size as independent risk factors for asthma, allergy, and eczema. Pediatr Allergy Immunol. 2003;14:464–9.
15. Wills-Karp M, Brandt D, Morrow AL. Understanding the origin of asthma and its relationship to breastfeeding. Adv Exp Med Biol. 2004;554:171–91.
16. Hedin K, Andre M, Molstad S, Rodhe N, Petersson C. Infections in families with small children: use of social insurance and healthcare. Scand J Prim Health Care. 2006;24:98–103.
17. Saunders NR, Tennis O, Jacobson S, Gans M, Dick PT. Parents' responses to symptoms of respiratory tract infection in their children. CMAJ. 2003;168:25–30.
18. Andre M, Hedin K, Hakansson A, Molstad S, Rodhe N, Petersson C. More physician consultations and antibiotic prescriptions in families with high concern about infectious illness–adequate response to infection-prone child or self-fulfilling prophecy? Fam Pract. 2007;24:302–7.
19. Magidson J, Vermunt JK. Chapter 10: latent class models. In: Kaplan D, editor. The sage handbook of quantitative methodology for the social sciences. London: Sage Publications Inc; 2004. p. 175–98.
20. Golding J, Pembrey M, Jones R. ALSPAC–the Avon Longitudinal Study of Parents and Children. I. Study methodology. Paediatr Perinat Epidemiol. 2001;15:74–87.
21. Falagas ME, Mourtzoukou EG, Vardakas KZ. Sex differences in the incidence and severity of respiratory tract infections. Respir Med. 2007;101:1845–63.
22. Simpson J, Smith A, Ansell P, Roman E. Childhood leukaemia and infectious exposure: a report from the United Kingdom Childhood Cancer Study (UKCCS). Eur J Cancer. 2007;43:2396–403.
23. Chang ET, Montgomery SM, Richiardi L, Ehlin A, Ekbom A, Lambe M. Number of siblings and risk of Hodgkin's lymphoma. Cancer Epidemiol Biomarkers Prev. 2004;13:1236–43.
24. Gibbs S, Surridge H, Adamson R, Cohen B, Bentham G, Reading R. Atopic dermatitis and the hygiene hypothesis: a case-control study. Int J Epidemiol. 2004;33:199–207.
25. Gergen PJ, Fowler JA, Maurer KR, Davis WW, Overpeck MD. The burden of environmental tobacco smoke exposure on the respiratory health of children 2 months through 5 years of age in the United States: Third National Health and Nutrition Examination Survey, 1988 to 1994. Pediatrics. 1998;101:E8.

26. Elliott L, Henderson J, Northstone K, Chiu GY, Dunson D, London SJ. Prospective study of breast-feeding in relation to wheeze, atopy, and bronchial hyperresponsiveness in the Avon Longitudinal Study of Parents and Children (ALSPAC). J Allergy Clin Immunol. 2008;122:49–54.
27. Bener A, Hoffmann GF, Afify Z, Rasul K, Tewfik I. Does prolonged breastfeeding reduce the risk for childhood leukemia and lymphomas? Minerva Pediatr. 2008;60:155–61.
28. Infante-Rivard C, Fortier I, Olson E. Markers of infection, breastfeeding and childhood acute lymphoblastic leukaemia. Br J Cancer. 2000;83:1559–64.
29. Oddy WH. A review of the effects of breastfeeding on respiratory infections, atopy, and childhood asthma. J Asthma. 2004;41:605–21.
30. Hagenaars JA, McCutcheon AL. Applied latent class analysis. Cambridge: Cambridge University Press; 2009.
31. Rabe-Hesketh S, Skrondal A. Classical latent variable models for medical research. Stat Methods Med Res. 2008;17:5–32.
32. Nylund KL, Asparoutiov T, Muthen BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Eq Modelling. 2007;14:535–69.
33. Muthen B, Muthen LK. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. Alcohol Clin Exp Res. 2000;24:882–91.
34. Vermunt JK, Magidson J. Latent class analysis. In: Lewis-Beck MS, Bryman A, Liao TF, editors. The sage encyclopedia of social science research methods. London: SAGE; 2003.
35. Muthén BO, Muthén LK. Mplus User's Guide. Los Angeles, CA: Muthén & Muthén; 2007.
36. Clark S, Muthen B. Relating latent class analysis results to variables not included in the analysis. 2009. [cited 18/07/09] Available from: http://www.statmodel.com/download/relatinglca.pdf.
37. von Linstow ML, Hogh M, Nordbo SA, Eugen-Olsen J, Koch A, Hogh B. A community study of clinical traits and risk factors for human metapneumovirus and respiratory syncytial virus infection during the first year of life. Eur J Pediatr. 2008;167:1125–33.
38. Caracciolo S, Minini C, Colombrita D, Rossi D, Miglietti N, Vettore E, Caruso A, Fiorentini S. Human metapneumovirus infection in young children hospitalized with acute respiratory tract disease: virologic and clinical features. Pediatr Infect Dis J. 2008;27:406–12.
39. Carroll KN, Gebretsadik T, Griffin MR, Dupont WD, Mitchel EF, Wu P, Enriquez R, Hartert TV. Maternal asthma and maternal smoking are associated with increased risk of bronchiolitis during infancy. Pediatrics. 2007;119:1104–12.
40. Duijts L, Ramadhani MK, Moll HA. Breastfeeding protects against infectious diseases during infancy in industrialized countries. A systematic review. Matern Child Nutr. 2009;5:199–210.
41. Quigley MA, Cumberland P, Cowden JM, Rodrigues LC. How protective is breast feeding against diarrhoeal disease in infants in 1990s England? A case-control study. Arch Dis Child. 2006;91:245–50.
42. Drevenstedt GL, Crimmins EM, Vasunilashorn S, Finch CE. The rise and fall of excess male infant mortality. Proc Natl Acad Sci USA. 2008;105:5016–21.
43. Mage DT, Donner EM. The fifty percent male excess of infant respiratory mortality. Acta Paediatr. 2004;93:1210–5.
44. DiFranza JR, Aligne CA, Weitzman M. Prenatal and postnatal environmental tobacco smoke exposure and children's health. Pediatrics. 2004;113:1007–15.
45. Ozmert EN, Kilic M, Yurdakok K. Environmental tobacco smoke: is it a risk factor for diarrhea in 6–18 months old infants? Cent Eur J Public Health. 2008;16:85–6.
46. Shenassa ED, Brown MJ. Maternal smoking and infantile gastrointestinal dysregulation: the case of colic. Pediatrics. 2004;114:e497–505.
47. Bianchi SM, Robinson JP, Milkie MA. Changing rhythms of American family life. New York: Russell Sage Foundation Publications; 2006.
48. McKinney PA, Alexander FE, Nicholson C, Cartwright RA, Carrette J. Mothers' reports of childhood vaccinations and infections and their concordance with general practitioner records. J Public Health Med. 1991;13:13–22.
49. Reboussin BA, Ip EH, Wolfson M. Locally dependent latent class models with covariates: an application to under-age drinking in the USA. J R Stat Soc Ser A Stat Soc. 2008;171:877–97.
50. McLachlan G, Peel D. Finite mixture models. New York: John Wiley & Sons; 2000.